BOREAL ENVIRONMENT RESEARCH 13: 359–369
ISSN 1239-6095 (print) ISSN 1797-2469 (online)

Helsinki 28 August 2008

# Influence of taxonomic resolution and data transformation on biotic matrix concordance and assemblage–environment relationships in stream macroinvertebrates

Jani Heino

*Finnish Environment Institute, Research Programme for Biodiversity, P.O. Box 413, FI-90014 University of Oulu, Finland (e-mail: jani.heino@ymparisto.fi)*

*Received 25 Apr. 2007, accepted 27 June 2007 (Editor in charge of this article: Johanna Mattila)*

Heino, J. 2008: Influence of taxonomic resolution and data transformation on biotic matrix concordance and assemblage–environment relationships in stream macroinvertebrates. *Boreal Env. Res.* 13: 359–369.

I examined patterns in stream macroinvertebrate assemblage structure and assemblage–environment relationships at four taxonomic levels (i.e. species, genus, family, and order) and based on four data types (i.e. presence/absence, logarithmic, square-root, and raw abundance) in a boreal drainage basin. Tests of resemblance matrix concordance within taxonomic levels showed that not all matrices were strongly correlated. Presence/absence matrices showed poor correlations with raw abundance matrices, possibly reflecting the fact that a few dominant species were responsible for assemblage patterns in the latter. By contrast, logarithmic and square-root data generally showed strong matrix correlations, and this pattern existed at the species, genus, and family levels. Assemblage–environment relationships were rather similar between species-, genus-, and family-level data sets, given that the same key environmental variables were included in the final environmental dissimilarity matrices in BIO-ENV analysis. In conclusion, researchers should pay considerable attention to data transformations when interpreting assemblage patterns at different taxonomic levels and comparing different studies, as different data transformations may provide differing information and lead to highly differing conclusions.

## Introduction

Many studies in ecology, conservation biology, and environmental assessment are based on the examination of whole ecological communities. Searches for patterns in community structure typically utilise data on the distribution of species across a set of sites and resulting species-by-sites matrix. Patterns in the species-by-sites matrix and derived resemblance matrices can be affected by the type of transformations used (e.g. Clarke and Warwick 2001). The raw abundance data can be reduced to either presence/absence form or transformed using various different transformations (e.g. logarithmic, square-root, and fourth-root). A typical approach in community analysis is to decrease the dominance of a few abundant species through transformation of species data (e.g. Clarke and Warwick 2001). However, although transformations are commonly used in community analyses, little attention has been directed to the fact that the biotic patterns and community–environment relationships may differ between different data types and transformations (e.g. Anderson *et al*. 2005). However, conclusions of different studies typically rely on

a single data type, including raw abundance data (e.g. Townsend *et al*. 1987), square-root transformed data (e.g. Johnson and Goedkoop 2002), logarithmically transformed abundance data (e.g. Heino *et al*. 2003), and presence/absence data (e.g. Malmqvist and Mäki 1994).

Community analyses are by no means restricted to the use of species-level data, but patterns at higher taxonomic levels may also provide interesting information about ecological assemblages. Many studies in conservation biology have examined concordance in taxon richness patterns between different taxonomic levels, and concluded that variability in species richness is often strongly correlated to that of higher taxon richness (e.g. Gaston and Williams 1993, Balmford *et al*. 2000). Less attention has been devoted to examining concordance between species-level and higher taxonomic-level assemblage patterns, i.e., two resemblance matrices between different taxonomic aggregations (e.g. Negi and Gadgil 2002). For example, Bowman and Bailey (1997) found that species- and genus-level resemblance matrices showed high correlations to family-, order- and even class-level matrices in freshwater macroinvertebrates. Marshall *et al*. (2006) also found that species-level matrices were strongly correlated with genus- and family-level matrices, and significant correlations existed even between species- and order-level matrices in stream macroinvertebrates. Furthermore, it has been found for marine macroinvertebrates that resemblance matrices constructed from species-, genus-, and family-level data are strongly correlated only if they are based on the same data transformation (e.g. Olsgard *et al*. 1997).

An important aspect of ecological community analysis is also the degree to which different types of data sets and taxonomic levels are related to environmental characteristics. For instance, if more abundant species show stronger relationships to the environment than less abundant species, then abundance data or transformed abundance data may be more easily explicable than presence/absence data by environmental characteristics. By contrast, if less common species are as important as (or more important than) abundant species for community–environment relationships, presence/absence data may provide well enough match to the environmental characteristics. There may similarly be differences in the responses of different taxonomic levels to environmental conditions (e.g. Olsgard *et al*. 1998). For instance, species-level data may be required to find a good match between variation in community structure and environmental conditions (e.g. Hawkins *et al*. 2000), as it is at the level of species and their traits where the responses to the environmental conditions occur (e.g. Poff 1997). However, the distributions of species may bear a high degree of randomness. In such cases, data on higher taxonomic levels may balance the noise in the distribution of species, as species belonging to, say, the same genus may show rather similar responses to the environment, and their combined occurrence covers the whole range of suitable environmental conditions for each species. Even higher taxa may show relatively strong relationships to environmental characteristics, and thus the examination of assemblage–environment relationships is also feasible at such higher taxonomic levels (e.g. Murphy and Davy-Bowker 2006). Furthermore, it has previously been shown that the assemblage–environment relationships may be rather invariant at species, genus and family levels in stream macroinvertebrates (Furse *et al*. 1984, Marchant *et al*. 1995, Feio *et al*. 2006).

Stream macroinvertebrates are typical objects of community ecological and environmental monitoring studies (Rosenberg and Resh 1993, Allan 1995), so there exists a large information base on their assemblage–environment relationships. For example, it has been shown repeatedly that a set of a few key environmental variables is needed to account for a moderate part of variability in community structure. These variables typically include stream size-related factors, acidity, nutrient concentrations, and water colour (Townsend *et al*. 1983, Wright *et al*. 1984, Malmqvist and Mäki 1994, Mykrä *et al*. 2007). The mentioned studies generally analysed patterns at the lowest possible taxonomic level with presence/absence or logarithmically transformed abundance data. By contrast, few previous studies had simultaneously examined the effects of both taxonomic resolution and data transformation on assemblage patterns and assemblage–environment relationships in stream macroinvertebrates, although such studies had

been conducted for marine macroinvertebrates (Olsgard *et al*. 1998, Clarke and Warwick 2001).

I examined the patterns of stream macroinvertebrate assemblage structure and assemblage–environment relationships at four taxonomic levels (i.e. species, genus, family, and order). Based on earlier studies, I expected that data sets with the same transformation (i.e. presence/absence, logarithmic transformation, square-root transformation, and raw abundance) would show stronger matrix concordance than matrices of different data transformation, irrespective of the taxonomic level. In other words, I thus expected that concordance among taxonomic levels would be strong if the matrices were based on the same data transformation. Furthermore, I assumed that the assemblage–environment relationships do not vary appreciably between different taxonomic levels, but either decreases or increases in the strength of the match between assemblages and the environment could be found in relation to the data transformations examined. I tested these assumptions by analysing a high-resolution data set collected from a set of near-pristine streams in northeastern Finland.

## Methods

### Study area

The study area is located in the Koutajoki drainage basin in northeastern Finland (centred on 66°20´N, 29°20´E; area extent of sampling sites is 2173 km$^2$), just south of the Arctic Circle. The bedrock of the study area is highly variable, with extensive occurrence of calcareous rocks. Accompanied by considerable relative altitudinal differences, this is mirrored in highly variable vegetation, ranging from old-growth coniferous forests to riparian deciduous woodlands and from nutrient-poor bogs to luxurious fens. These factors also provide the basis for a high variability of stream habitats across the region. Headwater streams and small rivers in the study area are characterised by circumneutral to alkaline water, low turbidity, and nutrient concentrations ranging from low to moderate (Table 1). The 34 stream sites surveyed for the present study represent typical headwater streams and small rivers in the region, and they are scattered across the three major tributaries of the Koutajoki drainage system: Oulankajoki, Kitkajoki and Kuusinkijoki.

### Benthic macroinvertebrates and explanatory variables

Macroinvertebrates were sampled during baseflow conditions in September 2002. At each site, the field crew took a two-minute kick-net (net mesh size 0.3 mm) sample covering most microhabitats present in a riffle of approximately 100 m$^2$. The two-minute sample was divided among the most important visible microhabitats in a riffle. This sampling effort typically yields more than 70% of species occurring at a site in a given season, mainly missing species that occur only sporadically in streams (Mykrä *et al*. 2006). Macroinvertebrates and associated material were immediately preserved in ethanol in the field, and they were taken to the laboratory for further processing and identification. Macroinvertebrates, including non-biting midges (Diptera: Chironomidae), were identified to the lowest taxonomic levels possible (i.e. species or genus) using available identification keys for northern Europe (e.g. Nilsson 1996, 1997, and references therein). All macroinvertebrates col-

**Table 1**. Mean, SE and range (Min. and Max.) of selected environmental variables and richness of different taxonomic levels across the study streams.

| Variables | Mean | SE | Min. | Max. |
|---|---|---|---|---|
| Depth (cm) | 17.06 | 0.88 | 7.33 | 27.07 |
| Current velocity (cm s$^{-1}$) | 25 | 2.01 | 9.17 | 56.77 |
| Macrophyte cover (%) | 36.61 | 5.42 | 0 | 99 |
| Particle size (see text) | 6.25 | 0.25 | 2.1 | 8.2 |
| Shading (%) | 21.93 | 3 | 1 | 69.1 |
| Stream width (m) | 3.22 | 0.48 | 0.6 | 15 |
| Conductivity (mS m$^{-1}$) | 8.98 | 0.94 | 2.53 | 24 |
| pH | 7.57 | 0.05 | 7 | 8.1 |
| $NO_2 + NO_3$ (µg l$^{-1}$) | 14.74 | 4.36 | 5 | 140 |
| Total P (µg l$^{-1}$) | 13.78 | 2.06 | 4 | 44 |
| Colour (mg Pt l$^{-1}$) | 66.35 | 8.97 | 10 | 210 |
| Species richness | 38.06 | 1.61 | 21 | 60 |
| Genus richness | 32.62 | 1.24 | 18 | 51 |
| Family richness | 22.79 | 0.73 | 13 | 31 |
| Order richness | 7.15 | 0.27 | 4 | 12 |

lected were identified, so no subsampling was conducted. This was done to ensure that even locally uncommon species were included in the analyses.

After macroinvertebrate sampling, the field crew measured several riparian and in-stream variables at each site. Shading by riparian trees was measured as percentage canopy cover at 20 locations in evenly-spaced cross-channel transects. Depth and current velocity (at 0.6 × depth) were measured at 40 random locations in cross-channel transects. Macrophyte cover and substratum particle size were assessed in ten 50 × 50 cm squares placed randomly in each riffle. The following classification of particle sizes (modified Wentworth scale; e.g. Allan 1995) was used: 0 = organic matter, 1 = sand (diameter 0.25–2 mm), 2 = fine gravel (2–6 mm), 3 = coarse gravel (6–16 mm), 4 = small pebble (16–32 mm), 5 = large pebble (32–64 mm), 6 = small cobble (64–128 mm), 7 = large cobble (128–256 mm), 8 = small boulder (256–400 mm), and 9 = large boulder (> 400 mm). The proportion of each particle size class was estimated for each square, and these estimates were subsequently averaged to give the mean substratum particle size for a site. Based on the above measurements, coefficients of variation were also calculated for depth, current velocity, macrophyte cover, and particle size to describe habitat heterogeneity. Mean stream width was also measured at each sampling site, based on five cross-channel measurements. Water samples were collected simultaneously with physical measurements, and they were subsequently analysed for pH, conductivity, $NO_2 + NO_3$, total phosphorus, and colour following Finnish national standards.

## Data types and statistical methods

Taxa-by-sites matrices were constructed for each of the four taxonomic levels, i.e., species, genera, families, and orders. Each taxonomic data set was further divided into four data types and transformations: presence/absence, logarithmically transformed abundance [$\log(x + 1)$], square-root transformed abundance, and raw abundance data. Thus, there were 16 biotic matrices for statistical analyses. All statistical analyses pertained to the routines in PRIMER ver. 6 (Clarke and Gorley 2006). Bray-Curtis similarities between all pairs of sites were first calculated from each taxa-by-sites matrix. This similarity coefficient is generally deemed to be highly suitable for analyses of quantitative assemblage data, for example, because (i) it neglects the double zero problem (Legendre and Legendre 1998), (ii) it ranges from 0 for totally dissimilar assemblages to 100 for totally similar assemblages, (iii) and a change of measurement unit does not affect its value (Clarke and Warwick 2001). For a comparison of the performance of the Bray-Curtis coefficient in relation to other similarity and distance coefficients, *see* Faith *et al*. (1987). This coefficient can also be used for presence/absence data, and then it is equal to Sørensen's coefficient (Clarke and Warwick 2001).

I first used the BIO-ENV analysis to examine the relationships between assemblage patterns of each biotic resemblance matrix and environmental variables. BIO-ENV works by relating biotic resemblance matrices to environmental distance matrices by calculating rank correlation between the two matrices analysed (Clarke and Warwick 2001). In practice, Bray-Curtis resemblance matrices were correlated with an Euclidean distance matrix of appropriately transformed and normalised (centred on respective mean and standardised by standard deviation) environmental variables. The situation of correlating two dissimilarity or distance matrix pertains to Mantel test type analyses (Legendre and Legendre 1998). The drawback of the original Mantel test is that one cannot easily evaluate the importance of different environmental variables in explaining patterns in biotic resemblance matrix. By contrast, BIO-ENV does this by searching for the best subset of environmental variables, leading to the strongest correlation between biotic resemblance matrix and the environmental distance matrix based on the reduced set of environmental variables. The strength of the correlation was assessed based on non-parametric Spearman's rank correlation ($r_s$) and the significance of the relationships was based on random permutations of the data (I used 99 permutations).

To compare assemblage patterns between different taxonomic levels and data transformations, I used another Mantel test type method

called RELATE (Clarke and Warwick 2001). This method can be used to compare the relationship between two different resemblance matrices by calculating Spearman's rank correlations. The significance of the relationships between two matrices is assessed based on a permutation test (I used 999 permutations). I admit that the tests between different taxonomic levels and data transformations are not independent; however, there is no other way to examine the concordance between such data matrices, and the same dependence problem is true for alternative methods such as the original Mantel test and Procrustes rotation (Legendre and Legendre 1998).

Finally, I used second-stage non-metric multidimensional scaling (NMDS) to further examine the concordance between biotic resemblance matrices (Clarke and Warwick 2001). In practice, second-stage analysis first produces a rank correlation matrix of resemblance matrices (in this case, the matrices of each transformation at the four taxonomic levels). The present analyses were based on Spearman's rank correlation. A NMDS ordination plot is then produced from the second-stage matrix, portraying the similarities and differences between original resemblance matrices. For ecological applications of second-stage NMDS, *see* Olsgard *et al*. (1998) and Bilton *et al*. (2006).

## Results

All taxa were retained in the statistical analyses, because the omission of rare taxa may reduce the ability to describe important assemblage gradients (Cao *et al*. 1998, Lenat and Resh 2001). There was a total of 175 species, 131 genera, 57 families, and 15 orders in the biotic data matrices. There was wide variability in taxon richness across the study sites. Species number varied from 21 to 60, genus number from 18 to 51, family number from 13 to 31, and order number from four to 12 (Table 1). There was thus much scope for variability in assemblage structure across sites irrespective of taxonomic level. Environmental conditions, particularly physical habitat variables, also showed considerable variation across sites (Table 1).

The BIO-ENV analysis showed that the relationships between assemblage structure and environmental variables were rather weak (Table 2).

**Table 2**. Summary of the BIO-ENV analysis of the relationships between biotic resemblance matrices and the best sets of environmental variables used to construct environmental dissimilarity matrices. Shown are Spearman rank correlations ($r_s$), significance, and the best subsets of environmental variables. Significant relationships ($P \leq 0.050$) are set in boldface. Environmental variables: 1 = depth; 2 = depth CV; 3 = current velocity; 4 = current velocity CV; 5 = macrophyte cover; 6 = macrophyte cover CV; 7 = Particle size; 8 = particle size CV; 9 = shading; 10 = stream width; 11 = conductivity; 12 = pH; 13 = $NO_2$ + $NO_3$; 14 = total phosphorus; 15 = water colour.

| Biotic matrix | $r_s$ | $P$ | Subset of environmental variables |
|---|---|---|---|
| Species presence/absence | 0.315 | 0.180 | 2, 4, 5, 6, 9, 10, 12, 13, 14 |
| Species log-transformation | 0.326 | 0.200 | 2, 5, 6, 9, 12 |
| Species square-root transformation | 0.338 | 0.090 | 2, 5, 6, 12 |
| Species raw abundance | 0.298 | 0.150 | 2, 5, 6, 12 |
| Genus presence/absence | **0.351** | **0.050** | 2, 4, 6, 8, 9, 10, 11, 12, 13, 15 |
| Genus log-transformation | 0.379 | 0.080 | 2, 4, 5, 6, 12 |
| Genus square-root transformation | **0.381** | **0.050** | 2, 5, 6, 12 |
| Genus raw abundance | 0.315 | 0.100 | 2, 5, 6, 12 |
| Family presence/absence | **0.376** | **0.050** | 2, 4, 5, 6, 8, 9, 10, 12 |
| Family log-transformation | **0.401** | **0.040** | 2, 5, 6, 7, 9, 12 |
| Family square-root transformation | **0.372** | **0.030** | 2, 5, 6, 12 |
| Family raw abundance | 0.274 | 0.250 | 2, 5, 6, 12 |
| Order presence/absence | 0.244 | 0.240 | 4, 6 |
| Order log-transformation | 0.358 | 0.110 | 2, 5, 6, 12 |
| Order square-root transformation | 0.264 | 0.370 | 2, 5, 12 |
| Order raw abundance | 0.236 | 0.360 | 5 |

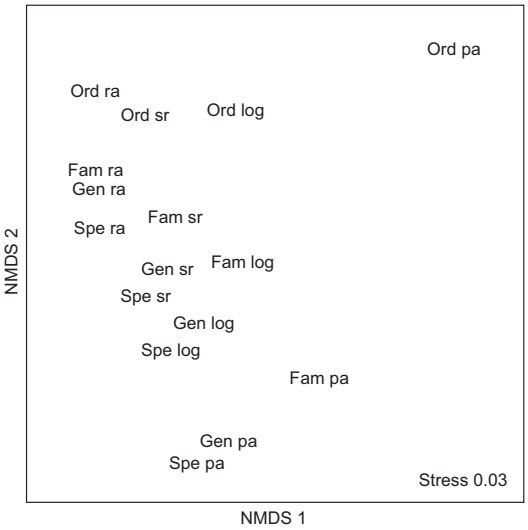


**Fig. 1**. An ordination plot from second-stage NMDS of different taxonomic levels and data types. Taxonomic levels: Spe = species; Gen = genus; Fam = family; Ord = Order. Data types: ra = raw abundance data; sr = square-root transformation; log = logarithmic transformation; pa = presence/absence data.

The best subsets of environmental variables were not significantly correlated with species-level assemblage patterns, with Spearman's rank correlations between biotic resemblance matrices and environmental distance matrices having $r_s$ values around 0.3. The situation was slightly different for genus-level data, with presence/absence and square-root transformed data showing significant relationships to the best subsets of environmental variables. Also, genus log-transformed and raw abundance data showed correlations with environmental variables that bordered significance. Spearman's $r_s$ values were slightly higher than those for species-level data. For family data, presence/absence, logarithmic, and square-root transformations were significantly correlated with environmental variables, with Spearman's $r_s$ values being around 0.4. Order-level resemblance matrices were not significantly correlated with environmental distance matrices (Table 2).

The best subsets of environmental variables had always a set of key variables influential at the species, genus, and family levels in BIO-ENV analyses (Table 2). These variables included coefficient of variation of depth, macrophyte cover, coefficient of variation of macrophyte cover, and pH. Further important variables were current velocity, shading, and stream width, but these were not as prevalent in the best subsets as the above mentioned variables. For order-level matrices, generally the same variables were incorporated in the environmental distance matrices as above, but the number of environmental variables in the best subsets was generally lower than that in the species-, genus-, and family-level analyses.

I considered $r_s = 0.9$ as the level of strong matrix concordance in RELATE, as it means that more than approximately 80% of the variability among resemblance matrices could be explained (Table 3). All pairwise comparisons of resemblance matrices were significant ($P < 0.01$), but due to non-independence, I concentrated only on the strength of the correlations. Spearman's rank correlations ($r_s$) between different biotic resemblance matrices showed that logarithmic and square-root transformed data were strongly correlated within taxonomic levels. For orders, however, logarithmic and square-root transformed data were not correlated at this threshold, although matrices based on square-root and raw abundance data were strongly correlated. A further important finding was that matrices at the adjacent taxonomic levels (e.g. species and genus) were strongly correlated if they were based on the same data transformation (e.g. presence/absence). In some cases, matrices based on logarithmic and square-root transformed data were also strongly correlated between taxonomic levels (e.g. species and genus).

Second-stage NMDS provided further insights into the relationships between different data sets (Fig. 1). Thus, order presence/absence data deviated sharply from all other data sets, including other order data sets, which were quite closely situated along the two dimensions in the NMDS plot. A general pattern shown by the remaining data sets was that species, genus, and family levels appeared to be quite closely situated if they were based on the same transformation, while different transformations of each taxonomic level were slightly further apart in the second-stage NMDS ordination plot. The second NMDS dimension seemed to group the data sets more clearly than the first dimension according to the transformation used.

**Table 3**. Spearman's rank correlation coefficients ($r_s$) between different biotic resemblance matrices from RELATE. Logarithmic and square-root transformations showed strong correlations within taxonomic levels. Matrices from adjacent taxonomic levels (e.g. species and genus) with the same data transformation (e.g. presence/absence) showed the strongest matrix correlations. In general, order level matrices showed weak correlations with other taxonomic levels, an exception being family raw abundance and order raw abundance.

| Matrix | Spa | Slo | Ssr | Sra | Gpa | Glo | Gsr | Gra | Fpa | Flo | Fsr | Fra | Opa | Olo | Osr |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Species presence/absence | – | | | | | | | | | | | | | | |
| Species log | 0.891 | – | | | | | | | | | | | | | |
| Species square-root | 0.794 | **0.961** | – | | | | | | | | | | | | |
| Species raw abundance | 0.560 | 0.787 | **0.918** | – | | | | | | | | | | | |
| Genus presence/absence | **0.931** | 0.851 | 0.760 | 0.545 | – | | | | | | | | | | |
| Genus log | 0.830 | **0.948** | **0.907** | 0.743 | 0.874 | – | | | | | | | | | |
| Genus square-root | 0.747 | **0.925** | **0.958** | 0.882 | 0.773 | **0.954** | – | | | | | | | | |
| Genus raw abundance | 0.505 | 0.736 | 0.863 | **0.947** | 0.512 | 0.731 | 0.891 | – | | | | | | | |
| Family presence/absence | 0.738 | 0.715 | 0.626 | 0.444 | 0.799 | 0.750 | 0.650 | 0.409 | – | | | | | | |
| Family log | 0.685 | 0.841 | 0.821 | 0.708 | 0.731 | **0.908** | 0.882 | 0.697 | 0.826 | – | | | | | |
| Family square-root | 0.643 | 0.846 | **0.908** | 0.887 | 0.666 | 0.877 | **0.953** | 0.898 | 0.663 | **0.922** | – | | | | |
| Family raw abundance | 0.467 | 0.688 | 0.825 | **0.930** | 0.468 | 0.677 | 0.847 | **0.980** | 0.383 | 0.672 | 0.892 | – | | | |
| Order presence/absence | 0.237 | 0.280 | 0.276 | 0.251 | 0.298 | 0.318 | 0.298 | 0.237 | 0.372 | 0.347 | 0.302 | 0.217 | – | | |
| Order log | 0.460 | 0.664 | 0.719 | 0.714 | 0.511 | 0.723 | 0.778 | 0.725 | 0.551 | 0.788 | 0.825 | 0.713 | 0.543 | – | |
| Order square-root | 0.467 | 0.677 | 0.790 | 0.858 | 0.481 | 0.681 | 0.818 | 0.890 | 0.423 | 0.689 | 0.868 | **0.909** | 0.314 | 0.857 | – |
| Order raw abundance | 0.408 | 0.603 | 0.733 | 0.844 | 0.402 | 0.579 | 0.743 | 0.888 | 0.305 | 0.588 | 0.787 | **0.921** | 0.200 | 0.693 | **0.954** |

## Discussion

Although is has been suggested that presence/absence data can be used to describe community patterns as efficiently as abundance data (e.g. Gauch 1982), there are a number of studies that have shown that data type and transformation do have effects on community patterns (Thorne *et al*. 1998, Clarke and Warwick 2001, Anderson *et al*. 2005). Thus, it was not surprising that second-stage NMDS and the tests of resemblance matrix concordance *within taxonomic levels* showed that not all matrices were strongly correlated. For instance, presence/absence matrices showed weak correlations with raw abundance matrices, possibly reflecting the fact that a few dominant species were responsible for assemblage patterns. By contrast, logarithmic and square-root transformed data generally showed strong matrix correlations ($r_s > 0.9$), and this pattern existed at the species, genus, and family levels. The likely reason for these patterns was that logarithmic and square-root transformations treated the abundance variation among taxa similarly by reducing the effects of abundant species on assemblage patterns, although the former transformation is generally considered to down-weight abundant taxa more severely than the latter (e.g. Clarke and Warwick 2001). Slightly lower correlations ($r_s$ = 0.8–0.9) were found between resemblance matrices based on presence/absence and logarithmically transformed abundance data at each taxonomic level, mirroring the fact that these data transformations produce fairly similar assemblage patterns by increasing the effects of less common species in comparison to untransformed abundance data.

Comparisons *between taxonomic levels* were variable in that the type of transformation affected the degree to which any two resemblance matrices were correlated. Similar results have emerged from studies of marine benthic invertebrates (Olsgard *et al*. 1997, Anderson *et al*. 2005). An important finding was that matrices at the adjacent taxonomic levels (e.g. species and genus) were strongly correlated if they were based on the same data transformation (e.g. logarithmic transformation). In some cases, matrices based on logarithmic and square-root data were also strongly correlated between taxonomic levels (e.g. species and genus). These findings are important given that higher taxonomic levels may be used to reproduce species-level assemblage patterns only if they are based on the same transformation of data. This finding is also in agreement with a number of similar studies, where resemblance matrices based on logarithmically transformed abundance data show strong correlation ($r > 0.9$) between different taxonomic levels (Marshall *et al*. 2006, Heino and Soininen 2007). A little weaker correlations ($r_s$ = 0.8–0.9) were detected between matrices of square-root transformed and raw abundance data between taxonomic levels, which likely portrayed the fact that, after square-root transformation, abundant taxa still have strong effects on assemblage patterns. In a similar comparison of resemblance matrices of different taxonomic levels, Bowman and Bailey (1997) found that the matrix correlations were stronger when quantitative abundance data were analysed than when presence/absence data were used. This finding is not entirely in agreement with the present ones, as even matrices based on presence/absence data of species and genera were strongly correlated, with the correlations being among the highest ones recorded in this study.

It has been suggested that the among-taxonomic level concordance in assemblage patterns is related to the ratios between species and higher taxa, with correlations becoming lower with increasing numbers of species in genera and genera in families in the region under study (Bowman and Bailey 1997, Hawkins *et al*. 2000). Hawkins and Norris (2000) suggested that when species richness is high, genera and families have undergone adaptive radiation, with species typically showing different environmental responses within families. This latter assumption is certainly true for the macroinvertebrate fauna of Finnish streams, as aquatic insect species within genera may have differing ecological niches and relationships to environmental factors (e.g. Heino 2005). Thus, it is not necessary that stream macroinvertebrate faunas in regions with rather low species diversity (e.g. boreal regions) show differing relative variability in environmental tolerances and optima than stream macroinvertebrates in regions with higher diversity (e.g. more southerly regions). The reason for

low regional species diversity in boreal regions is most likely related to fact that freshwater systems in these regions have developed since the latest ice age within 10 000 years or so (e.g. Brown and Lomolino 1998). Thus, the macroinvertebrate faunas in boreal regions are rather impoverished in comparison with regions that were not glaciated during the latest ice age or that were closer to potential refugia. Thus, boreal regions may still be under colonisation, and there may be a wide variety of stream environmental conditions not inhabited by stream macroinvertebrate species. Although there certainly are differences in the environmental preferences of stream macroinvertebrate species, it is highly likely that the responses to the environment are more similar between species of the same genus than between species of different genera. Thus, at least in low diversity regions, species-level assemblage patterns and assemblage–environment relationships are mirrored by patterns at the genus and family levels (Furse *et al*. 1984, Heino and Soininen 2007).

The assemblage–environment relationships were qualitatively similar between species-, genus-, and family-level data sets, given that the same key subset of environmental variables was included in the final environmental dissimilarity matrices. This finding is in accordance with a few studies conducted in other regions, and these studies have typically concluded that different taxonomic levels respond similarly to the same key environmental gradients, including stream size-related factors and acidity (Furse *et al*. 1984, Marchant *et al*. 1995, Waite *et al*. 2004). These are the same gradients that are known to be important for macroinvertebrate assemblages in boreal streams (Malmqvist and Mäki 1994, Heino *et al*. 2003, Johnson *et al*. 2004, Mykrä *et al*. 2007). In the present study, the most consistently important environmental factors that were related to biotic resemblance matrices were coefficient of variation of depth, macrophyte cover, coefficient of variation of macrophyte cover, and pH. These patterns imply that also habitat variability and habitat structure are important drivers of stream macroinvertebrate assemblages, at least in the absence of strong gradients in water chemistry. The similar relationships between biotic resemblance matrices and environmental distance matrices further suggest that the degree of correlation between taxonomic levels is, at least in part, due to similar environmental responses of different taxonomic levels. However, it must be stressed that the correlations between biotic resemblance matrices and environmental distance matrices were not significant in most cases, which may be due to some important, yet unmeasured environmental gradients (e.g. catchment characteristics).

Only genus- and family-level resemblance matrices based on presence/absence, logarithmic, and square-root transformed data were significantly correlated with environmental distance matrices. This finding raises the question why genus- and family-level data sets were more strongly correlated with environmental conditions than species-level data sets, although the absolute differences in the strength of these correlation were subtle. However, these findings might seem unexpected, because it is at the level of species where the responses to the environment should be most prevalent, as different species typically have different optima along environmental gradients (e.g. Bailey *et al*. 2001). There is at least one reason why one might also expect the opposite that genera and families be more strongly correlated to environmental conditions than species. For instance, one might envisage that genus- and family-level data do not include the noise inherent in complex and sometimes random distributions of species across sites (e.g. Bowman and Bailey 1997). Thus, given that some species may sometimes be absent from environmentally suitable sites due to temporary extinctions (e.g. Pulliam 2000), environmental relationships of species data may thus be weaker than those of genera and families. The distribution of genera or families naturally covers the whole range of environmental conditions inhabited by single species, and higher taxonomic levels may thus balance the sometimes sporadic characteristics of species-level information.

A number of implications for community ecology, conservation, and environmental assessment of streams emerged from the present analyses. First, genera and families may be used as adequate surrogates of species-level assemblage patterns in stream macroinvertebrate studies due to the relatively strong concordance

between biotic resemblance matrices. Second, if one intends to compare assemblage patterns at different taxonomic levels, then the comparisons should be based on the same transformation of data. Third, the correlations between biotic resemblance matrices were very strong only between the same data transformations at the adjacent taxonomic levels (e.g. species and genus), while the correlations weakened between taxonomic levels further apart (e.g. species and family). Finally, the assemblage–environment relationships may vary slightly among taxonomic levels, and according to the present results, genus and family assemblage structure could be most efficiently explained by environmental characteristics. The degree to which this latter finding holds in other regions with stronger environmental gradients in water chemistry remains to be studied rigorously. In conclusion, researchers should pay considerable attention to data transformations when interpreting community patterns, as different data transformations may provide differing information and, in the worst case, lead to highly differing conclusions. However, it is the very purpose of a study that eventually determines the taxonomic level (Bailey *et al*. 2001) and transformation to give answers to the questions posed (Clarke and Warwick 2001). For instance, although higher taxonomic levels may well reproduce species-level assemblage patterns and be suitable for rapid biodiversity surveys in guiding conservation planning and in environmental assessment, species-level data is obviously required for directing the conservation of rare species (Wright *et al*. 1996).

*Acknowledgements*: Several persons contributed to field sampling, sample processing, and identification of the macroinvertebrate samples. Foremost among these persons are H. Mykrä, T. Muotka and S. Koljonen. I warmly thank L. Paasivirta for identifying chironomid midges. Oulanka Research Station provided logistical support during the study. This study was financially supported by the Academy of Finland (grants to J. Heino and T. Muotka).

## References

Allan J.D. 1995. *Stream ecology. The structure and function of running waters*. Chapman and Hall, London, UK.

Anderson M.J., Connel, S.D., Gillanders B.M., Diebel, C.E., Blom, W.M., Saunders J.E. & Landers T.J. 2005. Relationships between taxonomic resolution and spatial scales of multivariate variation. *J. Anim. Ecol*. 74: 636–646.

Bailey R.C., Norris R.H. & Reynoldson T.B. 2001. Taxonomic resolution of benthic macroinvertebrate communities in bioassessments. *J. N. Amer. Benthol. Soc*. 20: 280–286.

Balmford A., Lyon J.E. & Lang R.M. 2000. Testing the higher-taxon approach to conservation planning in a megadiverse group: the macrofungi. *Biol. Cons*. 93: 209–217.

Bilton D.T., McAbendroth L., Bedford A. & Ramsay P.M. 2006. How wide to cast the net? Cross-taxon congruence of species richness, community similarity and indicator taxa in ponds. *Freshwat. Biol*. 51: 578–590.

Bowman M.F. & Bailey R.C. 1997. Does taxonomic resolution affect the multivariate description of the structure of freshwater benthic macroinvertebrate communities? *Can. J. Fish. Aquat. Sci*. 54: 1802–1807.

Brown J.H. & Lomolino M.V. 1998. *Biogeography*. Sinauer. Sunderland, USA.

Cao Y., Williams D.D. & Williams N.E. 1998. How important are rare species in aquatic community ecology and bioassessment? *Limnol. Oceanogr*. 43: 140–1409.

Clarke K.R. & Warwick R.M. 2001. *Change in marine communities. An approach to statistical analysis and interpretation*. PRIMER-E. Plymouth, UK.

Clarke K.R. & Gorley R.N. 2006. *PRIMER v6. User manual and tutorial*. PRIMER-E. Plymouth, UK.

Faith D.P., Minchin P.R. & Belbin L. 1987. Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio* 69: 57–68.

Feio M.J., Reynoldson T.B. & Graca M.A.S. 2006. The influence of taxonomic level on the performance of a predictive model for water quality assessment. *Can. J. Fish. Aquat. Sci*. 63: 367–376.

Furse M.T., Moss D., Wright J.F. & Armitage P.D. 1984. The influence of seasonal and taxonomic factors on the ordination and classification of running water sites in Great Britain and on the prediction of their macroinvertebrate communities. *Freshwat. Biol*. 14: 257–280.

Gaston K.J. & Williams P.H. 1993. Mapping the world's species — the higher taxon approach. *Biodiv. Lett*. 1: 2–8.

Gauch H.G. 1982. *Multivariate analysis in community ecology*. Cambridge University Press. Cambridge, UK.

Hawkins C.P. & Norris R.H. 2000. Effects of taxonomic resolution and use of subsets of the fauna on the performance of RIVPACS-type models. In: Wright J.F., Sutcliffe D.W. & Furse M.T. (eds.), *Assessing the biological quality of fresh waters: RIVPACS and other techniques*. Freshwater Biological Association, Ambleside, UK, pp. 217–228.

Hawkins C.P., Norris R.H., Hogue J.N. & Feminella J.W. 2000. Development and evaluation of predictive models for measuring the biological integrity of streams. *Ecol. Appl*. 10: 1456–1477.

Heino J. 2005. Positive relationship between regional distribution and local abundance in stream insects: a consequence of niche breadth or niche position? *Ecography* 28: 345–354.

Heino J., Muotka T., Mykrä H., Paavola R., Hämäläinen H. & Koskenniemi E. 2003. Defining macroinvertebrate assemblage types of headwater streams: implications for bioassessment and conservation. *Ecol. Appl.* 13: 842–852.

Heino J. & Soininen J. 2007. Are higher taxa adequate surrogates for species-level assemblage patterns and species richness in stream organisms? *Biol. Cons.* 137: 78–89.

Johnson R.K. & Goedkoop W. 2002. Littoral macroinvertebrate communities: spatial scale and ecological relationships. *Freshwat. Biol.* 47: 1840–1854.

Johnson R.K., Goedkoop W. & Sandin L. 2004. Spatial scale and ecological relationships between the macroinvertebrate communities of stony habitats of streams and lakes. *Freshwat. Biol.* 49: 1179–1194.

Legendre P. & Legendre L. 1998. *Numerical ecology.* Elsevier, Amsterdam, The Netherlands.

Lenat D.R. & Resh V.H. 2001. Taxonomy and stream ecology: The benefits of genus- and species-level identifications. *J. N. Amer. Benthol. Soc.* 20: 287–298.

Malmqvist B. & Mäki M. 1994. Benthic macroinvertebrate assemblages in north Swedish streams: environmental relationships. *Ecography* 17: 9–16.

Marchant R., Barmuta L.A. & Chessman B.C. 1995. Influence of sample quantification and taxonomic resolution on the ordination of macroinvertebrate communities from running waters in Victoria, Australia. *Mar. Freshwat. Res.* 46: 501–506.

Marshall J.C., Steward A.L. & Harch B.D. 2006. Taxonomic resolution and quantification of freshwater macroinvertebrate samples from an Australian dry-land river: the benefits and costs of using species abundance data. *Hydrobiologia* 572: 171–194.

Murphy J.F. & Davy-Bowker J. 2005. Spatial structure in lotic macroinvertebrate communities in England and Wales: relationships with physicochemical and anthropogenic stress variables. *Hydrobiologia* 534: 151–164.

Mykrä H., Heino J. & Muotka T. 2007. Scale-related patterns in the spatial and environmental components of stream macroinvertebrate assemblage variation. *Glob. Ecol. Biogeogr.* 16: 149–159.

Mykrä H., Ruokonen T. & Muotka T. 2006. The effect of sample duration on the efficiency of kick-sampling in two streams with contrasting substratum heterogeneity. *Ver Int. Ver. Theor. Ang. Limnol.* 29: 1351–1355.

Negi H.R. & Gadgil M. 2002. Cross-taxon surrogacy of biodiversity in the Indian Garhwal Himalaya. *Biol. Cons.* 105: 143–155.

Nilsson A.N. (ed.) 1996. *Aquatic insects of North Europe. A taxonomic handbook. Volume 1.* Apollo Books, Stenstrup, Denmark.

Nilsson A.N. (ed.) 1997. *Aquatic insects of North Europe. A taxonomic handbook. Volume 2.* Apollo Books, Stenstrup, Denmark.

Olsgard F., Somerfield P.J. & Carr M.R. 1997. Relationships between taxonomic resolution and data transformations in analyses of macrobenthic community along an established pollution gradient. *Mar. Ecol. Progr. Ser.* 149: 173–181.

Olsgard F., Somerfield P.J. & Carr M.R. 1998. Relationships between taxonomic resolution, macrobenthic community patterns and disturbance. *Mar. Ecol. Progr. Ser.* 172: 25–36.

Poff N.L. 1997. Landscape filters and species traits: towards mechanistic understanding and prediction in stream ecology. *J. N. Amer. Benthol. Soc.* 16: 391–409.

Pulliam R.H. 2000. On the relationship between niche and distribution. *Ecol. Lett.* 3: 349–361.

Rosenberg D.M. & Resh V.H. 1993. *Freshwater biomonitoring and benthic macroinvertebrates.* Chapman and Hall, New York, USA.

St. J. Thorne R., Williams W.P. & Cao Y. 1998. The influence of data transformation on biological monitoring studies using macroinvertebrates. *Wat. Res.* 33: 343–350.

Townsend C.R., Hildrew A.G. & Francis J. 1983. Community structure in some southern English streams: the influence of physicochemical factors. *Freshwat. Biol.* 13: 521–544.

Townsend C.R., Hildrew A.G. & Schofield K. 1987. Persistence of stream invertebrate communities in relation to environmental stability. *J. Anim. Ecol.* 56: 597–613.

Waite I.R., Herlihy A.L., Larsen D.P., Urquhart N.S. & Klemm D.J. 2004. The effects of macroinvertebrate taxonomic resolution in large landscape bioassessment: an example from the Mid-Atlantic Highlands, U.S.A. *Freshwat. Biol.* 49: 474–489.

Wright J.F., Moss D., Armitage P.D. & Furse M.T. 1984. A preliminary classification of running water sites in Great Britain based on macroinvertebrate species and prediction of community types using environmental data. *Freshwat. Biol.* 14: 221–256.

Wright J.F., Blackburn J.H., Gunn R.J.M., Furse M.T., Armitage P.D., Winder J.M. & Symes K.L. 1996. Macroinvertebrate frequency data for the RIVPACS III sites in Great Britain and their use in conservation evaluation. *Aquat. Cons. Mar. Freshwat. Ecosyst.* 6: 141–167.